Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 33 (2014) 253 – 260

Procedia Computer Science

CRIS 2014

The Quest for Research Information

Ina Blümel[a], Stefan Dietze[b], Lambert Heller[a], Robert Jäschke[b], Martin Mehlberg[a*]

[a]*German National Library of Science and Technology (TIB), Hannover*
[b]*L3S Research Center, Leibniz University Hannover*

Abstract

Research information, i.e., data about research projects, organisations, researchers or research outputs such as publications or patents, is spread across the web, usually residing in institutional and personal web pages or in semi-open databases and information systems. While there exists a wealth of unstructured information, structured data is limited and often exposed following proprietary or less-established schemas and interfaces. Therefore, a holistic and consistent view on research information across organisational and national boundaries is not feasible. On the other hand, web crawling and information extraction techniques have matured throughout the last decade, allowing for automated approaches of harvesting, extracting and consolidating research information into a more coherent knowledge graph. In this work, we give an overview of the current state of the art in research information sharing on the web and present initial ideas towards a more holistic approach for boot-strapping research information from available web sources.

Keywords: Research information; linked data; web crawling; information extraction

1. Introduction

Scholars at all levels – from PhD students to professors – need to be constantly updated about the actors involved in their subject area, the interactions between them and the publications they produce, as well as any breakthroughs achieved in their discipline and subject area (*Who analyses collective intelligence? In which projects have researchers worked (together)? Is there a noticeable trend concerning the allocation of computer science professorships? What third-party funding has been awarded? Which appointments have been made?*). In this context, it is becoming increasingly important to think outside the box, to conduct research in an interdisciplinary manner, and to establish networks with colleagues (*Which methods of complex network theory were already applied by theoretical physicists to my topic? Which universities have a professorship in "Web Science"?*). There is also a

* *E-mail addresses:* Ina.bluemel@tib.uni-hannover.de (Ina Blümel), dietze@L3S.de (Stefan Dietze), jaeschke@L3S.de (Robert Jäschke), Lambert.Heller@tib.uni-hannover.de (Lambert Heller), Martin.Mehlberg@tib.uni-hannover.de (Martin Mehlberg)

doi:10.1016/j.procs.2014.06.040

greater need for transferring knowledge to industry or the public sector (*Which industrial partners or public bodies are involved in e-learning projects?*). Further specific questions also arise in the process of writing project proposals, when researching state of the art or forming project consortia (*Which research projects exist in the field of simultaneous localisation? Which universities focus on the geosciences?*). Considerably more complex analyses may also be necessary, particularly in science research, on questions such as, e.g., *What is the structure of cooperative activities (publications/projects) between research institutions? Are there any differences between disciplines?* Other questions that could arise are: *What is the "typical" structure of a German university? How does the institute/staff structure of engineering institutes differ from that in natural science institutes?*

Therefore, researchers need to be well-informed at all times about the topics, collaborations, results, and trends in their discipline as well as adjacent ones. Such *research information* is needed from all scientific stakeholders, for research discovery and measurement, project planning and initialisation, evaluation and documentation of research activities. Stakeholders are researchers, research managers and administrators, policy makers, research councils and technology transfer organisations, the media and the general public (in the context of "citizen science"). Ideally, research information should be consistent, up to date and openly available. However, there is no holistic view of scientific "landscapes" concerning certain disciplines or countries: currently data is spread across institutional and personal websites, in research and project information systems, and in databases operated by commercial providers such as Elsevier, Google and Microsoft, where a range of varying schemas and interfaces are used. There are no substantial quantities of public data that can be analysed and processed by machines.

In addition, maintaining research information is also a burdensome task, causing most researchers unnecessary additional work. These days, researchers have to publish research results at the national and international levels, in different institutional contexts and in informal communities to ensure visibility and to facilitate the transfer of knowledge. This is done, e.g., by the maintenance and updating of web sites, research databases, publication management systems and presences on social media platforms. However, the simultaneous maintenance of these information channels has some undesirable consequences: Firstly, more and more resources are tied up in publication and exploration of information and secondly, the inconsistency of the databases that are spread over various platforms and web sites is growing. Consequently, by calling for the introduction of a "research core dataset" ("Kerndatensatz Forschung") [1], the German Council of Science and Humanities (Wissenschaftsrat) has acknowledged that the processing, linking and provisioning of research metadata are tasks of crucial importance.

In this work we review the state of publicly available research information (Sec. 2), identify current deficits (Sec. 3), and analyse existing approaches for capturing (Sec. 4), extracting (Sec. 5), representing (Sec, 6), and retrieving (Sec, 7) research information. We conclude with suggestions for the improvement of the situation in Sec, 8.

2. Sources for Research Information

At the moment (and presumably in the near future) an important primary source for research information are *web pages of institutions and researchers*, which are manually curated and often augmented with structured information, e.g., from literature databases. These websites fulfill an important role in academia: the public documentation and communication of any kind of research results. Since the sites are randomly distributed, heterogeneous, and typically unstructured, the information they contain can not be analysed without considerable effort. Therefore, a central collection and archival of structured data about research information is necessary.

Table 1 gives on overview on existing sources for research information. A traditional source are *literature databases* (A) which are established in many disciplines. They continue for research articles what libraries have been carrying out for more than 200 years for books. *Academic web search engines* (B) like Google Scholar (since 2004) often combine data from several digital libraries and other sources of structured data and augment it with heterogeneous data from the freely accessible web.

In favour of a considerably larger search space, the retrieval in search engines often is not as exact as in literature databases or library catalogues. However, meaningful profile pages[1] for researchers are generated from this broad

[1] e.g., http://scholar.google.de/citations?user=4ucFsi4AAAAJ

data foundation. Although the information content of such profiles exceeds pure publication lists, it represents an incomplete snapshot, since, e.g., patents, projects, or past activities and affiliations of researchers are not contained. The individually curated *websites of institutions and scientists* (C) as well as more recent crowd-sourcing approaches like *Wikipedia* (E) and the emerging commercial *social networks for researchers* (D) focus on the profile of the individual researcher including its temporal aspects – at the expense of even near- satisfactory analysis and retrieval capabilities over all profiles. Numerous – mostly proprietary and not cross-linked – implementations of *research information systems* (F) are currently established at an institutional level with research management departments as driving force (e.g. policy making, reporting, financial / project management). Over the past years, some countries and regions like Norway (CRIStin), Flanders (FRIS), the Netherlands (NARCIS), the Czech Republic (IS VaVaI) and Slovenia (SICRIS) managed to combine institutional and other data sources with the help of research information systems in such a way that an almost complete retrieval over temporally complete researcher profiles of the respective institution or country is possible. Experiences with research information systems from some European countries have been presented recently, [2] the situation and requirements in Germany have been described in [2].

Table 1. Sources for research information.

source	example	data source	data is reusable	integration of persistent indentifiers like ORCID implemented or planned	retrieval for specific disciplines or institutions is almost complete	data is international and/or cross-institutional	temporal aspects, like past engagements of researchers, are captured
A. literature database	PubMed, Scopus, DBLP, arXiv, CiteSeer	publisher, professional associations, etc.	*	*	yes	yes	**
B. search engine	Google Scholar, Microsoft Academic Search	web crawling (A, C, D, F)	no	no	*	yes	**
C. individual/ institutional website		manual input by researchers, often augmented by A	yes	no	no	no	yes
D. social network	Academia.edu, Mendeley, ResearchGate	manual input by researchers, often augmented by A	*	no	no	yes	*
E. Wikipedia/ Wikidata		manual input by Wikipedia authors	yes	yes	no	yes	yes
F. research information system	Atira Pure, Avedas Converis	master data of research facilities, often manually complemented by researchers and by A	*	yes	yes	no	yes
G. VIVO aggregator	VIVO Search, CTSA Search, AgriVIVO	pure aggregation from F	yes	yes	yes	yes	no

** In some (of the here mentioned) cases possible, but not always.*
*** Does not apply, because publication metadata is prevalent.*

A comprehensive *aggregation of research information* (G) is enabled by the open source tool VIVO [3] which aims for connecting research information within and across institutions (cf. Sec. 6). Originating in a predominantly US-based user group, it has lately been adopted by some European pilots and global networks like AgriVIVO. [3]

[2] http://www.forschungsinfo.de/kerndatensatz/en/index.php?mitteilungen
[3] http://www.agrivivo.net/

3. Deficits of the existing solutions

Despite the existence of the aforementioned sources, a comprehensive and exact overview on and analysis of activities of researchers is not possible in most countries and disciplines. So far, cross-institutional, comprehensive, reusable and structured research information can only be compiled where machine-readable data is available, thanks to ontologies such as VIVO, i.e., effectively only in life sciences in the English-speaking world.[4]

Web crawlers have been used for over ten years in academia to combine data from structured and unstructured sources. However, they are used at the expense of research information that goes beyond mere information about the publication, e.g., projects, completions, institutional memberships, awards, unconventional research products created by scientists. So far, web crawling has maximised the quantitative scope of crawled data in an unbalanced manner. In addition, commercial providers such as Google Scholar and Microsoft Academic Search restrict the reusability of their tool chain and search index. The datasets of most systems are not available under a free license, as is usually now the case with "open government data" and "linked library data", and their software is often proprietary.

In spite of these serious restrictions by the currently prevailing proprietary providers of transdisciplinary academic research engines their commercial practicability is in doubt: Elsevier, the third large provider besides Google and Microsoft, will stop its product "Scirus"[5] and the information science community has been speculating about a potential discontinuation of Google Scholar[6] since 2013. On the other hand, open source software has made considerable progress in web crawling and knowledge extraction in recent years. E.g., DBpedia [4] and YAGO [5] are semantically processing data extracted from Wikipedia and making it freely available in standardised, structured semantic web formats, enabling data and relationships to be analysed and machine processed easily. This has led to the creation of many applications and new research results that build on this knowledge, i.e., the "Linked Open Data (LOD) cloud" [6]. Other projects such as Common Crawl provide versioned crawls[7] with data from all kinds of websites. However, such approaches have not yet (or hardly) been applied to the domain of research information.

At present, *the research landscape is underrepresented in the Linked Open Data cloud, and there are no holistic approaches for the automated creation of scientific information based on Linked Open Data principles*. Machine-processable data about employees, developing young scholars, third-party funding and projects, research prizes and awards, patents, publications, etc. are not freely available to any appreciable degree. This is accompanied by the usually poorly developed possibilities for exploration and analysis: *The exploration of different facets and complex enquiries and analyses are not supported, changes to data over time are not documented or are not traceable*.

These problems are to be addressed, involving the development of a reusable infrastructure and relevant methods, that enable the creation, curation, exploration, and archiving of structured data about the research landscape. The tool pipeline to be developed combines *data capturing* and *information extraction and linking* (including data consolidation methods) in order to automatically generate Linked Open Data about the research landscape. The starting situation in the respective scientific areas is described in the following sections.

4. Data Capturing

Scientific data is often initially captured *manually*, e.g., when writing or submitting a publication or creating a web page. The effort for manual entry and curation is high, there are many redundancies (e.g., because no import is possible), and many systems do not provide any added value for researchers (e.g., re-using entered data for web pages). Manual capturing is successful when it is part of a process which is in the self-interest of the researchers, for example the submission of a research proposal (e.g., via CORDIS[8]) or a publication (e.g., on arXiv.org), or when automatically captured data merely needs to be complemented or corrected, as it is common for search engines and social networks. Moreover, these systems further benefit from crowdsourcing by allowing their users to add missing

[4] http://bit.ly/rXI4UZ
[5] http://www.scirus.com/
[6] http://bit.ly/13Y4IlL
[7] http://commoncrawl.org/
[8] http://cordis.europa.eu/

datasets. Such approaches are promising since the users need to (not have to!) correct or complement only few dataset which they typically know well. Thereby, they also retain a certain amount of control on their own data.

The benefit of *automatically* crawling the web to gather research data is that web pages are often primary data sources that are controlled by the affected entities (researchers, institutions) and that a large amount of data is freely available. This largely *unstructured* data is captured with *web crawlers* and subsequently structured data is extracted [7] – this is the standard procedure of current search engines. Several open source projects have provided usable and robust implementations of web crawlers, most notably Heritrix [8] and Apache Nutch.[9] The essential challenge is the *focused crawling* of relevant web pages [9] to enable an effective and efficient capturing of the desired data. Focusing can refer to the content or other properties of the web page (e.g., its file type or domain). A topical focus aims to crawl only pages related to a certain topic and is the matter of ongoing research [10]. Relevant in that context is the identification of pages of researchers using machine-learning approaches, as demonstrated in [11].

An alternative source to create a knowledge base of research information is *structured* data, for instance publication metadata that can be captured automatically using interfaces like OAI-PMH[10] or other implementations (e.g., ORCID is using the Scopus API[11]). Unfortunately, currently only a marginal amount of relevant data is available through such interfaces, which are also rarely standardized and often have access restrictions. Thus, the effort for implementation and maintenance of access to such repositories is high. Therefore, this approach is feasible only occasionally, e.g., to integrate subject-specific publication metadata or normalized person data from libraries.

5. Information Extraction and Linking

After crawling of unstructured information, structured data needs to be extracted. Given the lack of coherence of automatically extracted data, consolidation and linking of data is another important task, aiming to improve coherence and richness, also by taking into account background knowledge from related structured data sources.

Entity recognition is one of the major tasks within information extraction, and has been successfully applied in areas such as ontology generation, business intelligence, and text classification. It may encompass both named entity recognition (NER) and term extraction. NER selects and categorises entity names (such as persons, organisations, and location names), temporal expressions (dates and times), and certain types of numerical expressions (monetary values and percentages) in unstructured text. These may involve rule-based systems [12] or machine learning techniques [13]. Term extraction involves the identification and filtering of term candidates for the purpose of identifying domain-relevant terms or entities. The main aim in automatic term recognition is to determine whether a word or a sequence of words is a term that characterises the target domain. Most term extraction methods use a combination of linguistic filtering (e.g., possible sequences of part of speech tags) and statistical measures (e.g., TF-IDF) [15], to determine the salience of each term candidate for each document in the corpus [16].

Data consolidation has to cover a variety of areas such as enrichment, entity/identity resolution for disambiguation as well as clustering and correlation to consolidate disparate data. In addition, link prediction and discovery is of crucial importance to enable clustering and correlation of enriched data sources. A variety of methods for entity resolution have been proposed, using relationships among entities [17], string similarity metrics [18], as well as transformations [19]. An overview of the most important works in this area can be found in [20]. As opposed to entity correlation techniques, text clustering of documents exploits feature vectors, to represent documents according to contained terms [23]. Clustering algorithms measure the similarity across the documents and assign the documents to the appropriate clusters based on this similarity. In addition, web documents can be clustered based on their link structure [24], such as co-citations. Some web page clustering approaches use a combination of text and link structures [25]. Similarly, vector-based approaches have been used to map distinct ontologies and datasets [26].

As opposed to text clustering, entity correlation and clustering takes advantage of background knowledge from related datasets to correlate previously extracted entities. Therefore, link discovery is another crucial area to be

[9] http://nutch.apache.org/
[10] http://www.openarchives.org/pmh/
[11] http://searchapidocs.scopus.com/

considered. Graph summarization predicts links in annotated RDF graphs. Lehmann et al. [28] introduces a tool to discover relationships between objects in RDF datasets which was extended [29] for easy exploration and assessment of the graph. Seo et al. [32] proposed to use an RDF schema for finding relationships between two objects through their class relationships, resulting in a faster and more accurate approach. Related approaches are presented in [33] and [34], which explore the relation between two given objects from different ontologies.

Discussed general-purpose tools and methods are central to identifying semantically related entities and are the basis for specifically tailored approaches adapted to the specific case of research information.

6. Metadata for Research Information

There are lots of efforts being done to formalise research metadata, especially by determining standards and ontologies like the Common European Research Information Format (CERIF) [36], the CERIF-compliant FRAPO ontology for describing research project administrative information, or the Scholarly Contributions and Roles Ontology (SCoRO).[12] The Consortia Advancing Standards in Research Administration Information (CASRAI)[13] aims at ensuring interoperability of research information by the community-based establishment of best practices for data exchange and reuse and the development and maintenance of a common data dictionary. Another formalisation attempt is the VIVO ontology, building on widely used LOD namespaces (like, e.g., FOAF for describing agents like persons, or BIBO[14] for bibliographic resources like journal articles) and supporting links to persistent identifiers like ORCID. Yet, the VIVO core ontology is still tailored to the American research landscape. A mapping [37] between the VIVO ontology and the European CERIF will include most common CERIF enhancements.

7. Retrieval

The availability of an adequate data collection on the web does not guarantee the usability and utility for the satisfaction of the initially stated information needs. This requires at least appropriate query languages, although complex analyses often can only be performed offline on the whole dataset. In reality, however, the databases of most commercial providers are a business secret and can only be viewed manually on the web. Automatic capturing, e.g., via web crawling, is not allowed. On the other hand, access to some databases is possible (with restrictions) via application programming interfaces (e.g., for Microsoft Academic Search[15]) or as a download but only seldom are complete datasets available (e.g., for DBLP). Besides the OAI-PMH standard, which is common for repositories (e.g., PubMed and all DSpace-based systems), simple REST-based interfaces [38], SPARQL endpoints [39] as well as URL dereferencing are common. The latter provides resource-specific information via content negotiation using basic HTTP requests [40]. Such interfaces are important to enable linking of data bases in the web and to facilitate applications that provide value-added services. When data is accessible as a download, the most important aspect (besides the file format) is whether the complete data is available and whether primary data or extracted and further processed data is provided. Only a free and complete availability of all data in a well-documented format and schema (e.g., as for DBpedia) enables comprehensive analyses and reproducibility of scientific results.

The retrieval functionality of a system is tightly coupled with its database architecture, since how the data is stored influences how access is possible and which queries are feasible. Besides traditional relational databases, distributed approaches (e.g., HBase or CouchDB[16]) are common, in particular for big data. They promise a high scalability, fault tolerance and reliability. For Linked Open Data, in practice the following three approaches are relevant: (a) traditional RDF storage solutions like SESAME/OpenRDF, Jena or OWLIM,[17] (b) hybrid solutions like Virtuoso or the combination of relational databases for storage coupled with SPARQL interfaces and RDF mapping

[12] http://purl.org/cerif/frapo, http://purl.org/spar/scoro/
[13] http://casrai.org/
[14] http://vivoweb.org/ontology/core, http://xmlns.com/foaf/spec/, http://bibliontology.com/
[15] http://academic.research.microsoft.com/About/Help.htm#4
[16] https://hbase.apache.org/, https://couchdb.apache.org/
[17] http://www.openrdf.org/, https://jena.apache.org/, http://www.ontotext.com/owlim

approaches like D2R,[18] or (c) distributed triple stores like, e.g., Jena-HBase [41], HDRS[19] or other HBase-based solutions like H2RDF [42]. To category (b) belongs the currently from the Wikidata project [43] realised solution, which in particular stores the edit history. An extension of term-based search is facetted search which allows to further restrict search results according to various dimensions. Relevant and appropriate data models for the exchange of research information in the Linked Open Data cloud have been presented in Section 6.

8. Conclusion and Outlook

As motivated, the scientific community must be supported in the generation, maintenance and publication of data concerning research activities and results, and in the search for and exploration of such information.

Structuring research information according to Linked Open Data principles can enhance the discoverability of data, especially by adding useful context through harvested content from multiple research institutions and other sources. Assuming that in the future highly individualised, heterogeneous web pages of researchers and institutions will be important sources of information about their activities, we propose to develop solid, easy-to-use software tools for the extraction and provision of structured research information. Such software tools can be used to crawl the web sites of universities and research institutions and to extract metadata from them, which is then complemented and consolidated with freely available databases and authority files like VIAF or GND.[20] Outcomes can also be used for enhancing existing tools, e.g., the built-in VIVO Harvester.[21] Extracted research information can then be analysed, searched and made available as Linked Open Data in established formats such as VIVO and CERIF. Due to the nature of LOD, researchers using, e.g., VIVO interfaces can add meaningful objects and connections (including links to researchers and objects outside their own research institution) that can later be harvested and indexed both at the higher (discovery) level, as on the individual research institutions CRIS level, bootstrapping the building of (CERIF-based) CRIS. The provision of such data and tools would help to

- support scientists in devising new research issues and in the search for cooperation partners for projects,
- relieve the burden of maintaining research information,
- provide relevant information for allocations of third-party funding and for appointments,
- tap additional sources of alternative metrics for assessing the importance of scientific results and, ideally
- facilitate a transparent, evidence-based scientific policy.

The presented vision to dissolve current gaps and needs of exploring research information requires a joint effort of the relevant stakeholders to develop and implement the appropriate technology and curation mechanisms that ensure the free and enduring availability of research information.

References

1. Empfehlungen zu einem Kerndatensatz Forschung, Drucksache 2855-13, Wissenschaftsrat, Berlin, Germany (2013).
2. D. Beucke, A. Bliemeister, B. Ebert, E. Friedrichsen, L. Heller, S. Herwig, N. Jahn, M. Kreysing, D. Müller, M. Riechert, R. Tobias, Forschungsinformationssysteme in Hochschulen und Forschungseinrichtungen, Tech.rep., DINI AG Forschungsinformationssysteme (2014)..
3. D. B. Krafft, N. A. Cappadona, B. Caruso, J. Corson-Rikert, M. Devare, B. J. Lowe, V. Collaboration, VIVO: Enabling National networking of scientists, in: Proceedings of the Web Science Conference, Web Science Trust, 2010.
4. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia – a crystallization point for the web of data, Web Semantics 7 (3) (2009) 154–165.
5. F. M. Suchanek, G. Kasneci, G. Weikum, YAGO: a core of semantic knowledge, in: Proc. WWW, ACM, 2007, pp. 697–706.
6. C. Bizer, T. Heath, T. Berners-Lee, Linked data – the story so far, Int. Journal on SemanticWeb and Information Systems 5 (3) (2009) 1–22.
7. C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, New York, 2008.
8. G. Mohr, M. Kimpton, M. Stack, I. Ranitovic, Introduction to Heritrix, an archival quality web crawler, in: Proc. Int. Web Archiving Workshop, Bath, UK, 2004.

[18] http://virtuoso.openlinksw.com/, http://d2rq.org/d2r-server
[19] http://code.google.com/p/hdrs/
[20] http://viaf.org/, http://www.dnb.de/gnd
[21] https://wiki.duraspace.org/display/VIVO/VIVO+Harvester

9. S. Chakrabarti, M. van den Berg, B. Dom, Focused crawling: a new approach to topic-specific web resource discovery, Computer Networks 31 (11-16) (1999) 1623–1640.
10. J. Wu, P. Teregowda, J. P. F. Ramírez, P. Mitra, S. Zheng, C. L. Giles, The evolution of a crawling strategy for an academic document search engine: whitelists and blacklists, in: Proceedings of the 3rd Annual ACM Web Science Conference, ACM, 2012, pp. 340–343.
11. S. D. Gollapalli, C. Caragea, P. Mitra, C. L. Giles, Researcher homepage classification using unlabeled data, in: Proc. Int. Conf. on the World Wide Web, International World Wide Web Conferences Steering Committee, 2013, pp. 471–482.
12. D. Maynard, V. Tablan, C. Ursu, H. Cunningham, Y. Wilks, Named entity recognition from diverse text types, in: Recent Advances in Natural Language Processing, Tzigov Chark, Bulgaria, 2001.
13. Y. Li, K. Bontcheva, H. Cunningham, Adapting SVM for data sparseness and imbalance: A case study on information extraction, Natural Language Engineering 15 (2) (2009) 241–271.
14. G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manage. 24 (5) (1988) 513–523..
15. D. Maynard, Y. Li, W. Peters, NLP techniques for term extraction and ontology population, in: Proc. Conf. on Ontology Learning and Population: Bridging the Gap Between Text and Knowledge, IOS Press, Amsterdam, The Netherlands, 2008, pp. 107–127.
16. P. Deane, A nonparametric method for extraction of candidate phrasal terms, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, ACL, 2005, pp. 605–613.
17. X. Dong, A. Halevy, J. Madhavan, Reference reconciliation in complex information spaces, in: Proc. SIGMOD, , ACM, 2005, pp. 85–96.
18. W. W. Cohen, P. Ravikumar, S. E. Fienberg, A comparison of string distance metrics for name-matching tasks, in: Proceedings of the IJCAI-03 Workshop on Information Integration, 2003, pp. 73–78.
19. S. Tejada, C. A. Knoblock, S. Minton, Learning domain-independent string transformation weights for high accuracy object identification, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2002, pp. 350–359..
20. A. K. Elmagarmid, P. G. Ipeirotis, V. S. Verykios, Duplicate record detection: A survey, Trans. Knowl. and Data Engin, 19 (1) (2007) 1–16.
21. D. Boley, Principal direction divisive partitioning, Data Mining and Knowledge Discovery 2 (4) (1998) 325–344.
22. A. Z. Broder, S. C. Glassman, M. S. Manasse, G. Zweig, Syntactic clustering of the web, Computer Networks and ISDN Systems 29 (8-13) (1997) 1157–1166, papers from the Sixth International World Wide Web Conference.
23. A. Hotho, A. Maedche, S. Staab, Ontology-based text clustering, in: Proc. Workshop "Text Learning: Beyond Supervision" at IJCAI, 2001.
24. Y. Wang, M. Kitsuregawa, Use link-based clustering to improve web search results, in: Proc. WISE, IEEE, 2001, pp. 115–124.
25. Y. Wang, M. Kitsuregawa, Evaluating contents-link coupled web page clustering for web search results, in: Proceedings of the 11th International Conference on Information and Knowledge Management, CIKM '02, ACM, 2002, pp. 499–506.
26. S. Dietze, A. Gugliotta, J. Domingue, Exploiting metrics for similarity-based semantic web service discovery, in: Proceedings of the International Conference on Web Services, IEEE, 2009, pp. 327–334.
27. S. Dietze, J. Domingue, Exploiting conceptual spaces for ontology integration, in: Data Integration through Semantic Technology (DIST2008) Workshop at 3rd Asian Semantic Web Conference, 2008.
28. J. Lehmann, J. Schuppel, S. Auer, Discovering unknown connections – the DBpedia relationship finder, in: Proceedings of the 1st Conference on Social Semantic Web, Vol. 301 of CEUR Proceedings, Leipzig, Germany, 2007.
29. P. Heim, S. Lohmann, T. Stegemann, Interactive relationship discovery via the semantic web, in: Proc. ESWC, Springer, 2010, pp. 303–317.
30. P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, T. Stegemann, RelFinder: Revealing relationships in RDF knowledge bases, in: Semantic Multimedia, Vol. 5887 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, 2009, pp. 182–187.
31. S. Lohmann, P. Heim, T. Stegemann, J. Ziegler, The RelFinder user interface: Interactive exploration of relationships between objects of interest, in: Proceedings of the 15th International Conference on Intelligent User Interfaces, ACM, 2010, pp. 421–422.
32. D. Seo, H. Koo, S. Lee, P. Kim, H. Jung, W.-K. Sung, Efficient finding relationship between individuals in a mass ontology database, in: U- and E-Service, Science and Technology, Vol. 264 of Communications in Computer and Information Science, Springer, 2011, pp. 281–286.
33. M. Sabou, E. Motta, Relation discovery from the semantic web, in: Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference, Vol. 408 of CEUR Proceedings, 2008.
34. Y.-J. Han, S.-B. Park, S.-J. Lee, S. Park, K. Kim, Ranking entities similar to an entity for a given relationship, in: PRICAI 2010: Trends in Artificial Intelligence, Vol. 6230 of Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, 2010, pp. 409–420.
35. K. Anyanwu, A. Maduko, A. Sheth, SemRank: ranking complex relationship search results on the semantic web, in: Proceedings of the 14th International Conference on the World Wide Web, WWW '05, ACM, 2005, pp. 117–127.
36. B. Jörg, CERIF: The common european research information format model, Data Science Journal 9 (2010) CRIS24–CRIS31.
37. L. Lezcano, B. Jörg, M.-A. Sicilia, Modeling the context of scientific information: Mapping VIVO and CERIF, in: Advanced Information Systems Engineering Workshops, Springer, 2012, pp. 123–129.
38. R.T. Fielding, Architectural styles and the design of network-based software architectures, PhD thesis, University of California, Irvine (2000).
39. E. Prud'hommeaux, A. Seaborne, SPARQL query language for RDF, W3C (2008). URL http://www.w3.org/TR/rdf-sparql-query/
40. T. Heath, C. Bizer, Linked data: Evolving the web into a global data space, Synthesis Lectures on the SemanticWeb: Theory and Technology 1 (1) (2011) 1–136.
41. V. Khadilkar, M. Kantarcioglu, B. Thuraisingham, P. Castagna, Jena-HBase: A distributed, scalable and efficient RDF triple store, in: Proceedings of the ISWC 2012 Posters & Demonstrations Track, Vol. 914 of CEUR Proceedings, Boston, USA, 2012.
42. N. Papailiou, I. Konstantinou, D. Tsoumakos, N. Koziris, H2RDF: adaptive query processing on RDF data in the cloud, in: Proceedings of the 21st International Conference Companion on WWW, ACM, 2012, pp. 397–400.
43. D. Vrandečić, Wikidata: A new platform for collaborative data collection, in: Proc. WWW Companion, ACM, 2012, pp. 1063–1064.